<Bioinformatics Center> Pathway Engineering

CITATION:

<Bioinformatics Center> Pathway Engineering. ICR Annual Report 2007, 13: 64-65

ISSUE DATE:

2007-03

URL:

http://hdl.handle.net/2433/65517

Bioinformatics Center
- Pathway Engineering -

http://www.bic.kyoto-u.ac.jp/pathway/index.html

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
TAKIGAWA, Ichigaku
(D Eng)

PD
SHIGA, Motoki
(D Eng)

PD (JSPS)
WAN, Raymond
(Ph D)

PD (JSPS)
ZHU, Shanfeng
(Ph D)

Visitors

Dr ANGELOPOULOS, Nikolaos — University of Edinburgh, UK, 8–29 October 2006
Prof WONG, Limsoon — National University of Singapore, Singapore, 15 December 2006

Scope of Research

With the recent advancement of experimental techniques in molecular biology, research in modern life science is shifting to the comprehensive understanding of a biological mechanism consisting of a variety of molecules. Our focus is placed on molecular mechanisms in biological phenomena, represented by biological networks such as metabolic and signal transduction pathways. Our research objective is to develop techniques based on computer science and/or statistics to systematically understand biological entities at the cellular and organism level.

Research Activities (Year 2006)

Presentations

Gene Sequence Ranking Based on Expression Profiles for Metabolic Pathway Analysis, SigBio Meeting, Information Processing Society of Japan, Takigawa I and Mamitsuka H, Sapporo, Japan, 9 February.

A Probabilistic Model-based Approach for Biomedical Text Mining, Mamitsuka H, First Japan-Taiwan Bilateral Symposium on Bioinformatics, Tokyo, Japan, 14 March.

ProfilePSTMM: Capturing Tree-structure Motifs in Carbohydrate Sugar Chains, Aoki-Kinoshita K F, Ueda N, Mamitsuka H and Kanehisa M, Fourteenth International Conference on Intelligent Systems for Molecular Biology (ISMB 2006), Fortaleza, Brazil, 7 August.

A New Efficient Probabilistic Model for Mining Labeled Ordered Trees, Hashimoto K, Aoki-Kinoshita K F, Ueda N, Kanehisa M and Mamitsuka H, Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), Philadelphia, USA, 23 August.

Applying Gaussian Distribution-dependent Criteria to Decision Trees for High-Dimensional Microarray Data, Wan R, Takigawa I and Mamitsuka H, VLDB Workshop on Data Mining in Bioinformatics, Seoul, Korea, 11 September.

Machine Learning and Pattern Recognition Using Mathematica, Takigawa I, Japan Mathematica Users Group Workshop 2006, Kobe, Japan, 28 October.

Learning Probabilistic Models for Mining Labeled Ordered Trees, Mamtisuka H, Third Japanese-German Frontiers of Sciences Symposium, Heidelberg, Germany, 3 November.

Learning a Probabilistic Model for Labeled Ordered Trees, Mamitsuka H, Second Taiwan-Japan Bilateral Symposium on Bioinformatics, Tainan, Taiwan, 9 November.

Learning a Probabilistic Model for Labeled Ordered Trees, Mamitsuka H, Workshop on Scientific Computing: Models, Algorithm and Applications, Hong Kong, China, 8 December.

Grants

Mamitsuka H, Probabilistic Model-based Method for Mining from Structured Data in Bioinformatics, Research Grant from Okawa Foundation for Information and Telecommunications, 1 September 2005–31 August 2006.

Mamitsuka H, Developing a Parameter Estimation Method for Efficient Systems Biology Based-on Machine Learning, Research Grant from Kayamori Foundation of Information Science Advancement, 1 January 2006–31 December 2007.

Takigawa I, Large-Scale Biological Information Processing Based on Computational Geometric Structures and Adaptive Sampling, Grant-in-Aid for Young Scientist (B), 1 April 2006–31 March 2008.

A New Efficient Probabilistic Model for Mining Labeled Ordered Trees

Mining frequent patterns is a general and important issue in data mining. Complex and unstructured (or semi-structured) datasets have appeared in major data mining applications, including text mining, web mining and bio-informatics. Mining patterns from these datasets is the focus of many of the current data mining approaches. We focus on labeled ordered trees, typical datasets of semi-structured data in data mining, and propose a new probabilistic model and its efficient learning scheme for mining labeled ordered trees. The proposed approach significantly improves the time and space complexity of an existing probabilistic modeling for labeled ordered trees, while maintaining its expressive power. We evaluated the performance of the proposed model, comparing it with that of the existing model, using synthetic as well as real datasets from the field of glycobiology. Experimental results showed that the proposed model drastically reduced the computation time of the competing model, keeping the predictive power and avoiding overfitting to the training data. Finally, we assessed our results using the proposed model on real data from a variety of biological viewpoints, verifying known facts in glycobiology.

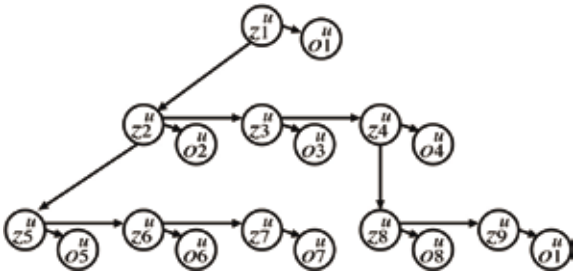

Figure 1. Graphical model of the proposed probabilistic model.

Applying Gaussian Distribution-Dependent Criteria to Decision Trees for High-Dimensional Microarray Data

Biological data presents unique problems for data analysis due to its high dimensions. Microarray data is one example of such data which has received much attention in recent years. Machine learning algorithms such as support vector machines (SVM) are ideal for microarray data due to its high classification accuracies. However, sometimes the information being sought is a list of genes which best separates the classes, and not a classification rate. Decision trees are one alternative which do not perform as well as SVMs, but their output is easily understood by non-specialists. A major obstacle with applying current decision tree implementations for high-dimensional datasets is their tendency to assign the same scores for multiple attributes. We then propose two distribution-dependant criteria for decision trees to improve their use fullness for microarray classification. We empirically demonstrated the advantage of the presented distributions using a variety of real microarray datasets as well as synthetic datasets.

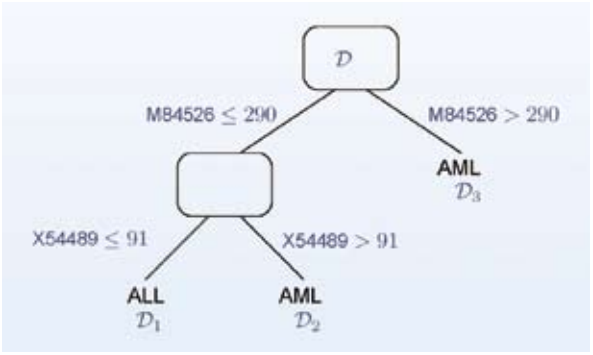

Figure 3. A typical example of a decision tree.

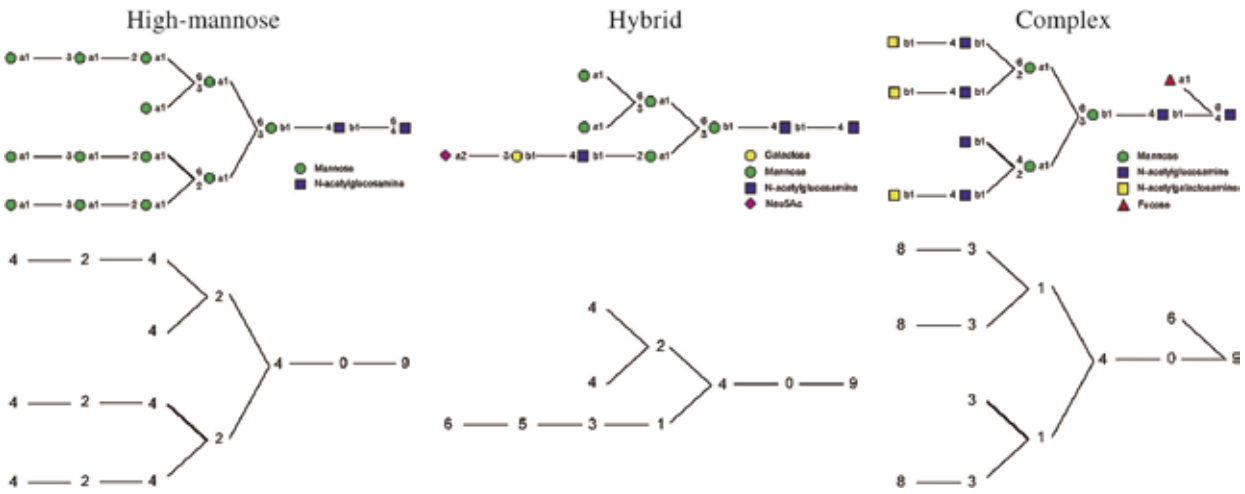

Figure 2. (top) The actual glycans, and (bottom) the most likely state paths.